Project 2
Ames Housing Price

Presented by:
Jerome,
Janet,
& Chee wee

Content

- Problem Statement
- Dataset Introduction
- Data Cleaning and EDA
- Preprocessing and modeling
- Evaluation and Conceptual understanding
- Conclusions and Recommendation

Introduction and Problem Statement

Target Audience

Sellers and buyers in the housing market

Problem Statements

Our main aim for this project is to utilize regression modelling to identify features that can be useful in predicting the housing sale price. The predictive modelling that we obtain for this project can be useful to house owners who are planning to sell their houses. They would be able to understand the features that have an influence on the value of their homes so that they will be able to better set the sale price for their own homes. For buyers of houses, this can assist them to understand better the value of the houses that they are buying by looking out for features that will influence the sale price of a house.

Dataset introduction

We will be using the following data set for the analysis on this project.

- Train.csv: Ames housing data set

In this data set, there are a total of 81 columns and 2051 rows. Out of the 81 columns, there are 80 explanatory variables describing the different aspects of residential homes in Ames, Iowa, and 1 column which contains the sale price of the houses. There are a total of 24 nominal, 23 ordinal, 14 discrete, and 20 continuous variables.

Dataset cleaning

Pool QC	2042
Misc Feature	1986
Alley	1911
Fence	1651
Fireplace Qu	1000
Lot Frontage	330
Garage Finish	114
Garage Qual	114
Garage Yr Blt	114
Garage Cond	114
Garage Type	113
Bsmt Exposure	58
BsmtFin Type 2	56
Bsmt Cond	55
Bsmt Qual	55
BsmtFin Type 1	55
Mas Vnr Area	22
Mas Vnr Type	22
Bsmt Full Bath	2
Bsmt Half Bath	2
Garage Area	1
Garage Cars	1
Total Bsmt SF	1
Bsmt Unf SF	1
BsmtFin SF 2	1
BsmtFin SF 1	1

- For Pool QC, we understand that NA value means that there is no pool. Therefore, for this column, we will substitute NA as None.
- For Misc Feature, we understand that NA value means that there are no miscellaneous features. Therefore, for this column, we will substitute NA as None.
- For Alley, we understand that NA value means that there is no alley access. Therefore, for this column, we will substitute NA as None.
- For Fence, we understand that NA value means that there is no fence. Therefore, for this column, we will substitute NA as None.
- For Fireplace Qu, we understand that NA value means that there are no fireplaces. Therefore, for this column, we will substitute NA as None.
- For Lot Frontage, we noted that there were quite a number of outliers based on our visualization, and therefore, we used the median as the value to populate for the missing values.
- For the columns relating to garage (i.e. Garage Finish, Garage Qual, Garage Yr Blt, Garage Cond, Garage Type), we understand that NA values means that there is no garage. However, we noted that for the column 'Garage Type', it seems that there is one less row which was not indicated as NA. Upon investigation, we noted that there was a row in which the garage type is indicated as Detached but all related columns on garage was missing. Therefore, we will drop this row of data from our analysis. In addition, for the row in which there was an odd year of 2207, as we are not able to make an inference from any columns to determine the Garage year built, we will drop this row from our analysis.

Dataset cleaning

Pool QC	2042
Misc Feature	1986
Alley	1911
Fence	1651
Fireplace Qu	1000
Lot Frontage	330
Garage Finish	114
Garage Qual	114
Garage Yr Blt	114
Garage Cond	114
Garage Type	113
Bsmt Exposure	58
BsmtFin Type 2	56
Bsmt Cond	55
Bsmt Qual	55
BsmtFin Type 1	55
Mas Vnr Area	22
Mas Vnr Type	22
Bsmt Full Bath	2
Bsmt Half Bath	2
Garage Area	1
Garage Cars	1
Total Bsmt SF	1
Bsmt Unf SF	1
BsmtFin SF 2	1
BsmtFin SF 1	1

- For Mas Vnr Type and Mas Vnr Area, we understand that NA value means that there are no Mas Vnr type. Therefore, for this column, we will substitute NA as None for Mas Vnr and substitute NA as 0 for Mas Vnr Area.

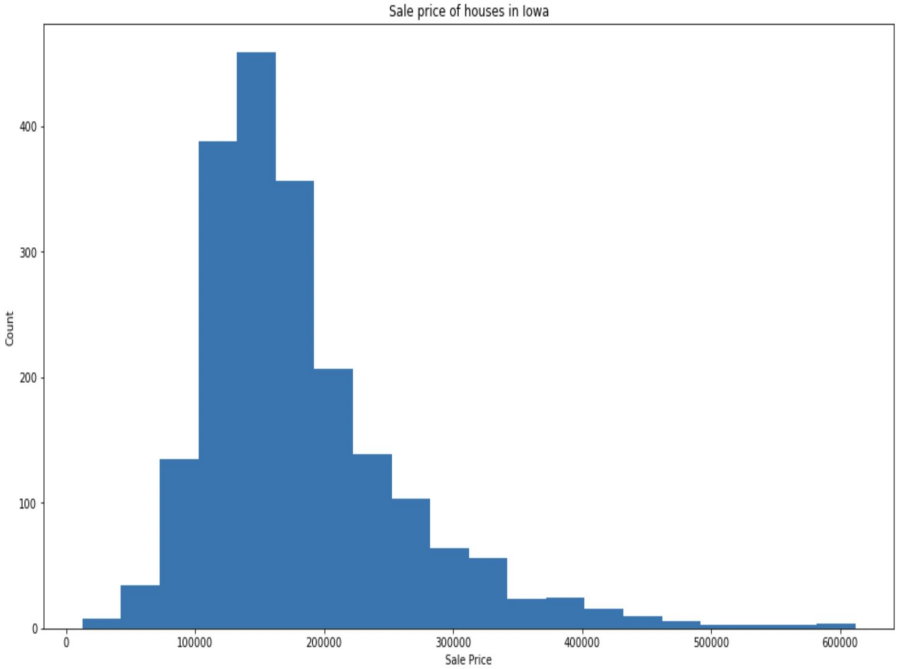
- For basement full bath and basement half bath columns, upon our investigation, we noted that the 2 rows in which NA is indicated for the column 'Basement Full Bath' are the same rows in which NA is indicated for the column 'Basement Half Bath'. As we understand that there are no basements for these 2 rows, we will replace NA with 0 accordingly.

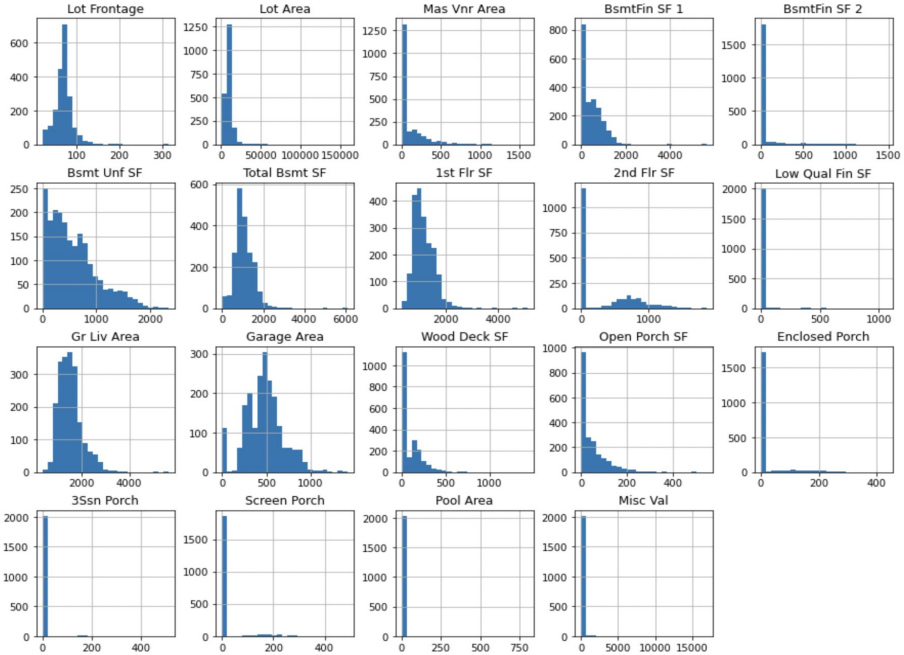
- For the columns relating to Bsmt Exposure, BsmtFin Type 2, Bsmt Cond, Bsmt Qual, BsmtFin Type 1, we noted that there were 55 rows which are NA for Bsmt Cond, Bsmt Qual, BsmtFin Type 1 which would indicate that there is no basement. However, for the other 2 columns, there are 58 and 56 rows which are NA respectively. Upon investigation, we noted that there were 3 rows in which the basement quality is indicated as Gd and the basement condition is indicated as TA but information on basement exposure is missing. Therefore, we will drop this 3 rows of data from our analysis.

- For the remaining columns relating to basement which have missing values, as there is no basement for this house, we will indicate and replace NA with 0 for the 'Total Bsmt SF', 'Bsmt Unf SF', 'BsmtFin SF 2' and 'BsmtFin SF 1' columns.

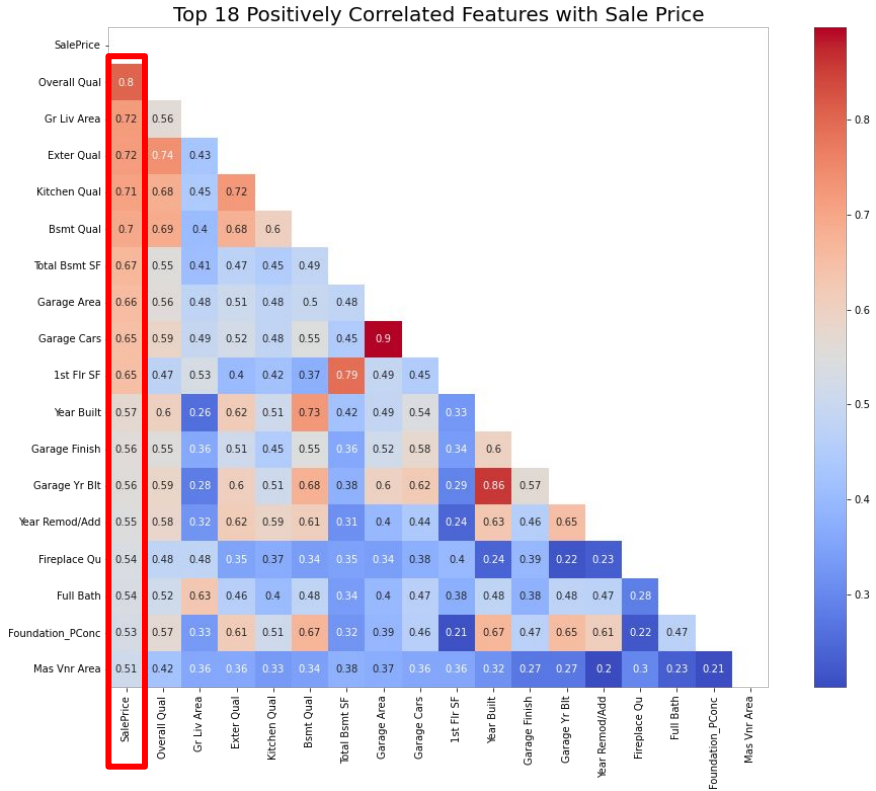
- Additionally, we conducted a cross check on the columns in which we have changed due to the missing values to ensure that it makes sense. During our cross check, we noted that there are 5 rows in the dataset in which Mas Vnr Type is indicated as None but the area was not 0. As we are not able to verify if these values are correct, we will be dropping these rows from our data set for the analysis.

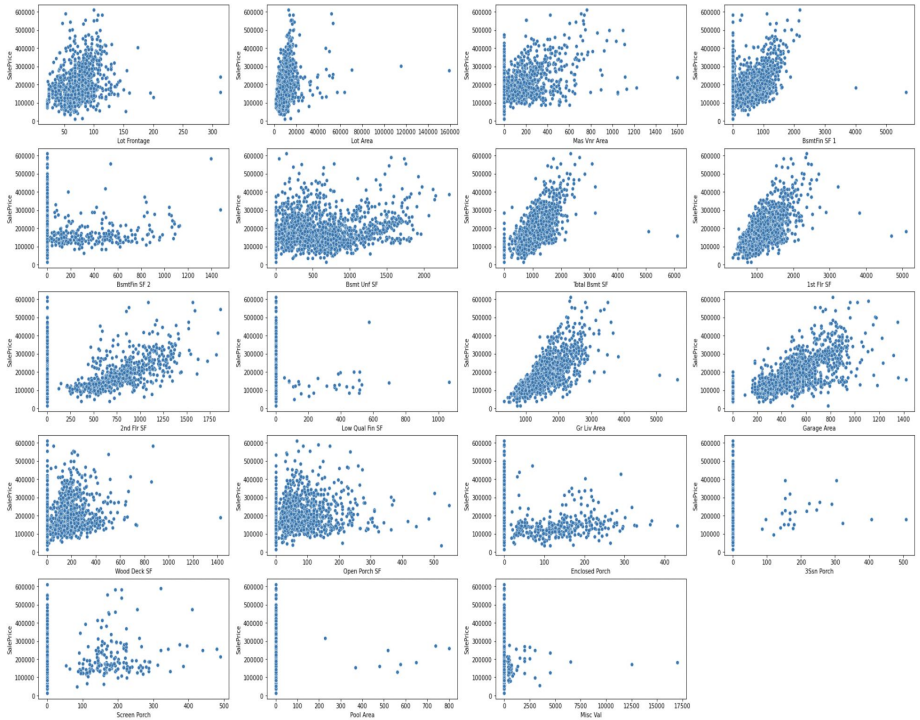
Exploratory Data analysis

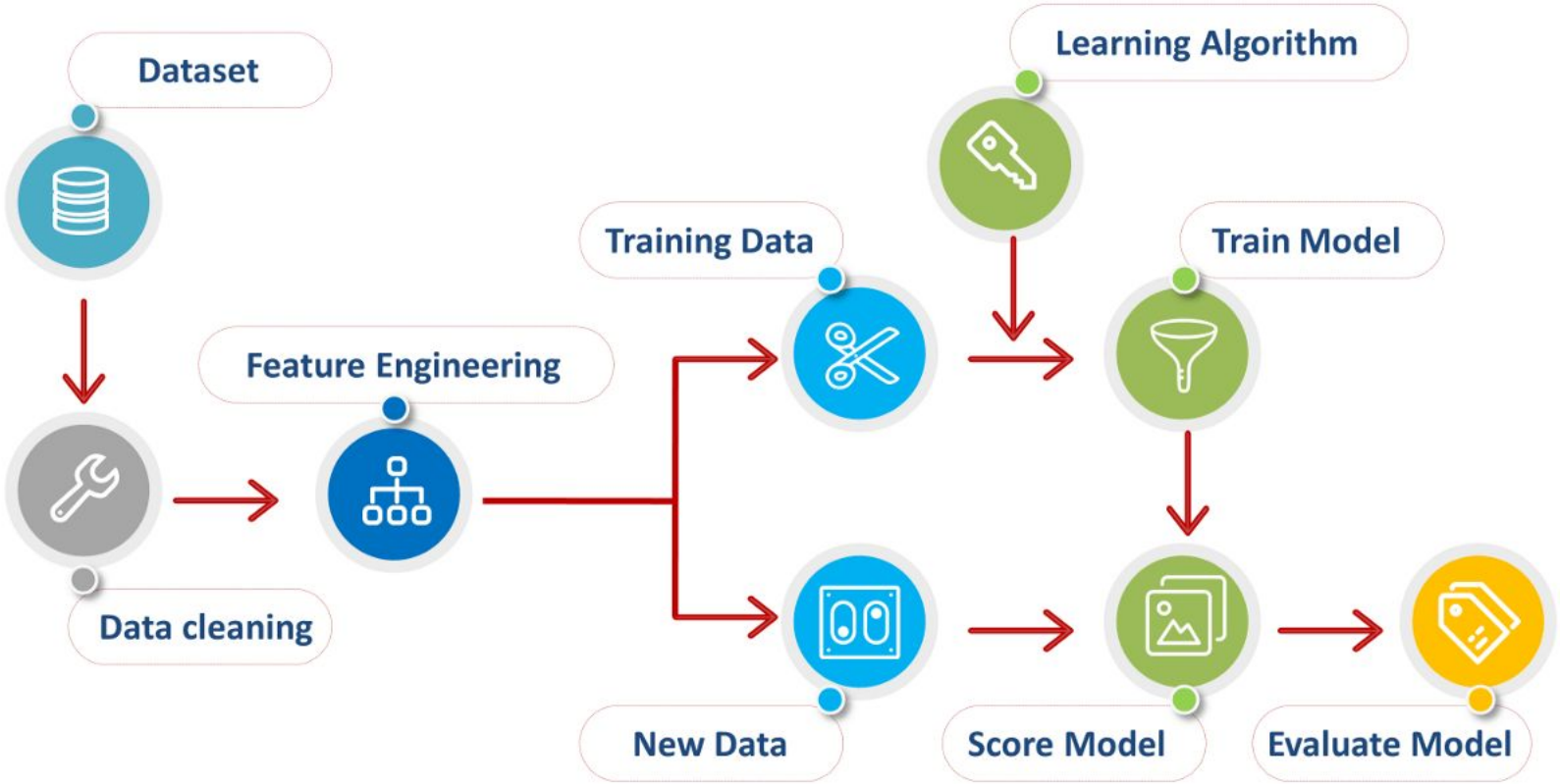




Exploratory Data Analysis






Dataset
Learning Algorithm
Training Data
Train Model
Feature Engineering
Data cleaning
New Data
Score Model
Evaluate Model

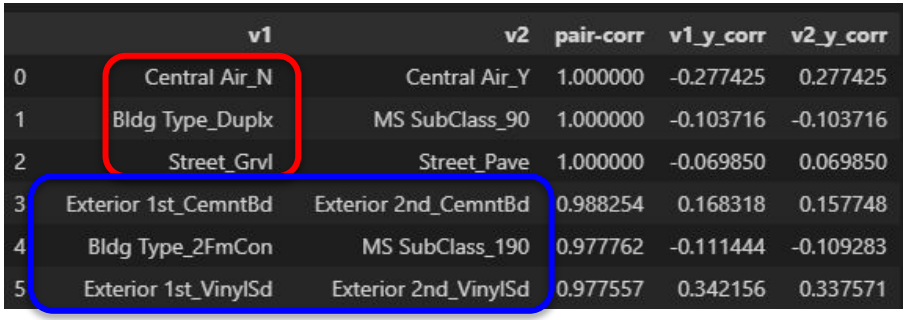
source

Dimensionality Reduction

- By dummifying, dataset shape increased from **80** to **228** features
  - Pros - increased dimensionality of data.
  - Cons - the risk of **overfitting** the model (noise features may be assigned non-zero coefficients due to chance associations with target variables)
- Dimensionality reduction techniques:
  - Correlation
  - Variance analysis
  - Recursive Feature Elimination

Feature Engineering - High Pairwise Correlation

|Var 1| Var 2| Pair-Corr | Var 1 vs Target | Var 2 vs Target



	v1	v2	pair-corr	v1_y_corr	v2_y_corr
0	Central Air_N	Central Air_Y	1.000000	-0.277425	0.277425
1	Bldg Type_Duplx	MS SubClass_90	1.000000	-0.103716	-0.103716
2	Street_Grvl	Street_Pave	1.000000	-0.069850	0.069850
3	Exterior 1st_CemntBd	Exterior 2nd_CemntBd	0.988254	0.168318	0.157748
4	Bldg Type_2FmCon	MS SubClass_190	0.977762	-0.111444	-0.109283
5	Exterior 1st_VinylSd	Exterior 2nd_VinylSd	0.977557	0.342156	0.337571

- Identify the High Pairwise Correlated Variables
- **Drop** or **Combine** Variables
  - Create interaction columns for Exterior features
  - Drop variables due to perfect pairwise correlation score of 1
  - Drop variables with lower absolute correlation to sale price
  - Combine variables (e.g. ‘Garage QualCond’)

Feature Engineering - Low Variance

Variables | var < 0.009

Sale Type_ConLD	0.008232
Neighborhood_Veenker	0.008232
Neighborhood_NPkVill	0.008232
MS SubClass_75	0.007751
Roof Matl_Tar&Grv	0.007271
House Style_2.5Unf	0.006789

- Many features are approximately constant, do not improve model performance, these violate the multivariate normality assumption of MLR
- **Combine** variables to make them statistically significant (e.g. ‘Sale type’, ‘Foundation’)
- **Drop 57** variables with variance below variance threshold of 0.009

Feature Engineering -
Recursive Feature Elimination

Overall Quality
Basement Full/ Half Bath
Total Bedrooms/ Kitchen/ Rooms
Fireplace
Garage Cars

- Recursive Feature Elimination (**RFE**) - fits a lr model, rank features based on data trained algorithm and removes the weakest feature(s) until the specified number of features is reached.
- **Drop** 28 variables
- Final # of features : 120

Summary of Feature Engineering

Type of variables	**Method**	**Example/ Results**
Ordinal Categorical Variables	Ordinal Encoding	13 variables remain
Nominal Categorical Variables	One Hot Encoding	10 variables remain
Interaction terms	Combine	Qual * Cond [Overall, Ext., Bsmt., Garage]

Modeling & Evaluation

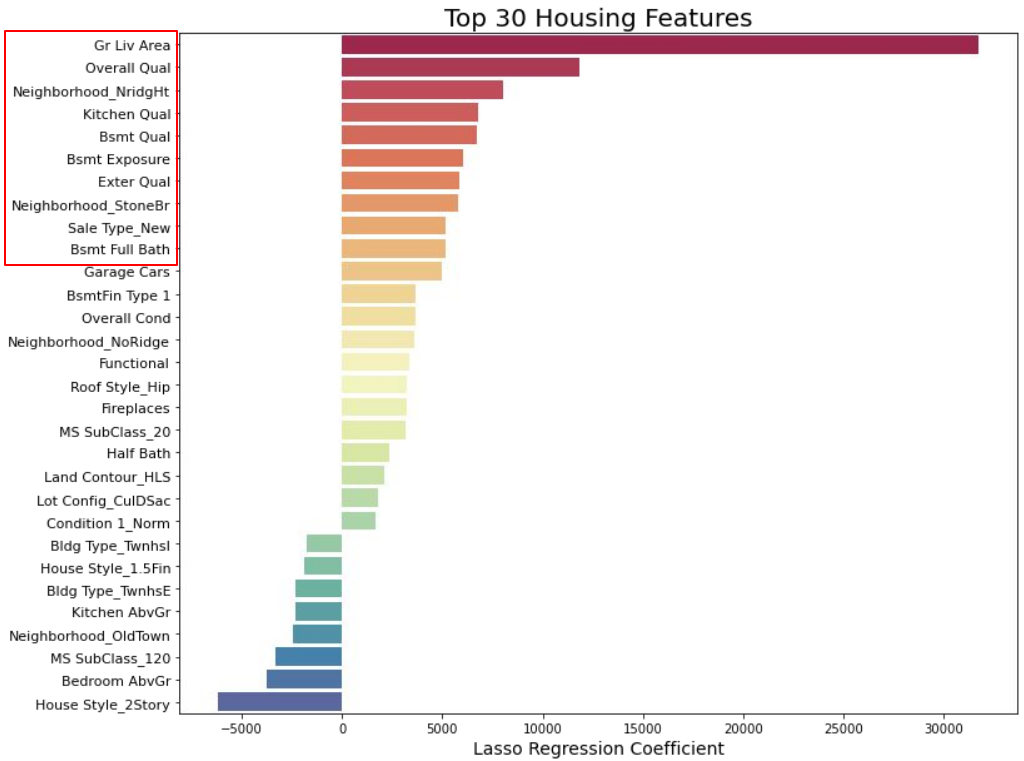
Model	CV_R2_score	Test_R2_score	Test_RMSE	Optimal_Alpha
LinearReg	0.888	0.906	24387	NA
Ridge	0.890	0.905	24564	43.212
Lasso	0.891	0.905	24556	294.367
ElasticNet	0.890	0.904	24610	0.065

- **Parameters:**
  - **Train-Test split : 75%-25%**
  - **StandardScaler**
- **Model is slightly overfitting, accuracy decreased between training dataset and CV dataset**

Modeling & Evaluation

- **LASSO Regression**
  - **R2 score: 0.91**
  - **RMSE score: $ 24,556**
- Strong linear relationship between predicted sale price and actual sale price

Top Housing Features



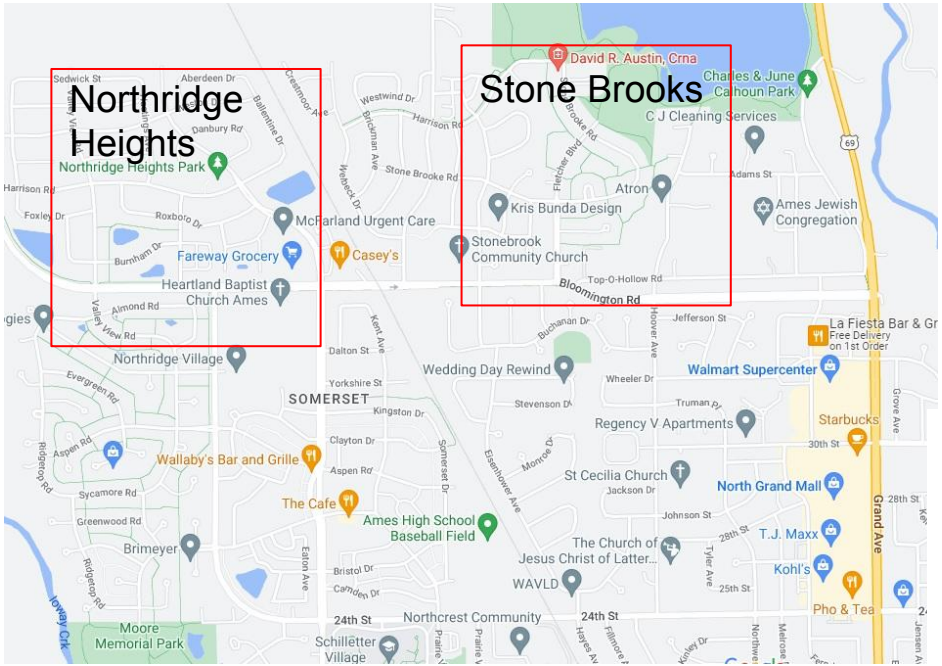
Interpreting the coefficients

Holding all else constant, for every unit increase in the Gr Liv Area, we expect an increase of $32000 to the SalePrice

Features that may add value to house price

1. Big house (Gr Liv Area),
2. Type of materials and finishing used and house condition (Overall Qual, Overal Cond)
3. Good Bsmt Exposure (garden, walkout basement)
4. New houses (Sale Type - New)
5. Neighbourbood Stone Brook, Northridge Heights

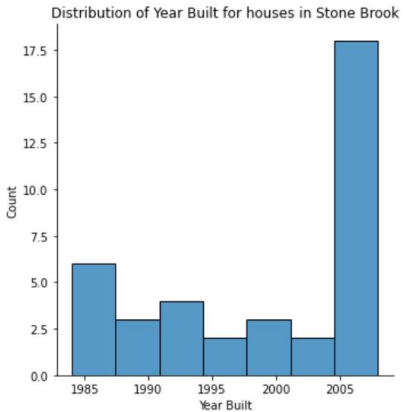
Houses in Stone Brook and Northridge Heights

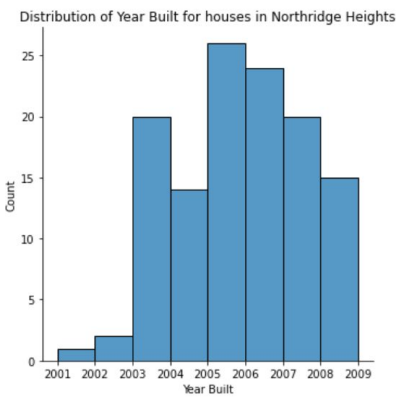


Observations:

1. Stone Brook and Northridge Heights are nearby to each other, and a short drive to nearby open spaces, parks, food, grocery, clinic, supermarket, shopping mall, school, churches
2. The average size of houses in Stone Brook (1980 sqft) and Northridge Heights(1940 sqft) are bigger when compared to the average size in Ames(1500 sqft)
3. The houses in Stone Brook and Northridge Heights are also newer, less than 10 years old from 2010

The observations are consistent with the features that may add value





Conclusion & Recommendation

Conclusion

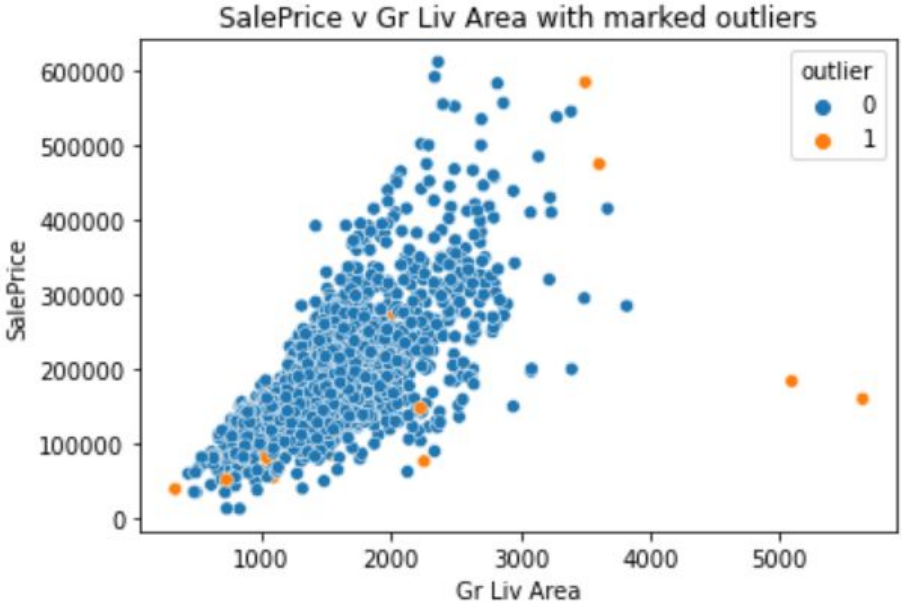
A production model is developed to predict the prices of houses in Ames, Iowa using R2 as the selection metric. The analysis has also found the following strong predictors of house prices

1. Big houses - Big living area (Gr Liv Area)
2. New houses - New sales (Sales type - New)
3. Open - garden/walkout basement (Bsmt Exposure)
4. Used very good materials and finishing and are in very good condition(Overall Qual, Overall Cond)
5. Neighbourhood that command a price premium are Stone Brook or Northridge Heights. They are close to each other and are a short drive to nearby amenities

Recommendations

- For house owners: Consider priortising renovations that will improve the materials and finishings of your house and the house condition as these factors are likely to improve the SalePrice
- For house buyers: Expect to pay a price premium for bigger, newer and open houses that used very good materials and finishing and are in very good condition. Good neighbourhood to invest are Stone Brook or Northridge Heights

Appendix - Outliers



Identify outliers using IsolationForest